UNIVERSITI PUTRA MALAYSIA

EFFICIENT ACCESS OF REPLICATED DATA IN DISTRIBUTED DATABASE SYSTEMS

MUSTAFA BIN MAT DERIS

EFFICIENT ACCESS OF REPLICATED DATA IN DISTRIBUTED DATABASE SYSTEMS

By

MUSTAFA BIN MAT DERIS

Thesis Submitted in Fulfillment of the Requirement for the Doctor of Philosophy in the Faculty of Computer Science and Information Technology University Putra Malaysia

September 2001

Dedicated to my beloved mother Chik Bt Omar and father Mat Deris bin Muda

"Thank you for your support"

Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Doctor of Philisophy

EFFICIENT ACCESS OF REPLICATED DATA IN DISTRIBUTED DATABASE SYSTEMS

By

MUSTAFA MAT DERIS

September 2001

Chairman : Assoc. Prof. Dr. Ali Mamat

Faculty : Computer Science and Information Technology

Replication is a useful technique for distributed database systems where a data object will be accessed (i.e., read and written) from multiple locations such as from a local area network environment or geographically distributed world wide. This technique is used to provide high availability, fault tolerance, and enhanced performance.

This research addresses the performance of data replication protocol in terms of data availability and communication costs. Specifically, this thesis present a new protocol called Three Dimensional Grid Structure (TDGS) protocol, to manage data replication in distributed database systems (DDS). The TDGS protocol is based on the logical structure of sites/servers in order to form a read or a write

quorum in the DDS. The protocol provide high availability for read and write operations with limited fault-tolerance at low communication cost. With TDGS protocol, a read operation is limited to two data copies, while a write operation is required with minimal number of copies. In comparison to other protocols, TDGS requires lower communication cost for an operation, while providing higher data availability.

A system for building reliable computing over TDGS Remote Procedure (TDGS-RP) system has also been described in this research. The system combines the replication and transaction techniques and embeds these techniques into the TDGS-RP system. The model describes the models for replicas, TDGS-RP, transactions, and the algorithms for managing transactions, and replicas.

CAPAIAN BERKESAN BAGI DATA REPLIKA DI DALAM SISTEM PANGKALAN DATA TERAGIH

Oleh

MUSTAFA MAT DERIS

September 2001

Chairman : Prof. Madya Dr. Ali Mamat

Faculty : Sains Komputer dan Teknologi Maklumat

Replikasi merupakan teknik yang penting bagi system pangkalan data teragih di mana data objek dicapai (iaitu baca atau tulis) dari beberapa lokasi seperti dari rangkaian setempat atau mana-mana tempat diseluruh dunia. Teknik ini digunakan untuk menyediakan ketersediaan yang tinggi, toleransi-kesalahan, dan peningkatan prestasi.

Tesis ini memaparkan prestasi protocol replikasi data dalam bentuk ketersediaan data dan kos komunikasi. Tesis ini mempersembahkan protokol baru dipanggil protokol Struktur Grid Berdimensi Tiga (TDGS) untuk mengurus replikasi data di dalam system pangkalan data teragih (DDS). Protokol ini berdasarkan kepada struktur logical pelanggan/tempat untuk membentuk korum baca atau tulis dalam DDS. Protokol ini menyediakan ketersediaan yang tinggi dengan kos komunikasi rendah. Dengan protokol TDGS, operasi baca memerlukan hanya dua salinan data, sementara bagi operasi tulis memerlukan jumlah salinan yang minima.

Dibandingkan dengan protokol-protokol lain, protokol TDGS memerlukan kos komunikasi rendah, dan menyediakan ketersediaan data yang tinggi.

Satu system untuk membangunkan pengkomputeraan yang dipercayai ke atas system TDGS-RP juga dijelaskan. Sistem ini menggabungkan teknik replikasi dan transaksi, dan menggunakan teknik ini ke dalam system TDGS-RP. Ia menjelaskan model bagi replica, TDGS-RP, transaksi, dan algorithma untuk mengurus transaksi dan replica.

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude and deep appreciation to the committee chairman, Assoc. Prof. Dr. Ali Mamat for his contribution, guidance, ideas, and time towards my thesis. I would also like to thank the committee members, assoc. Prof. Dr. Md. Yazid Mohd Saman and Dr. Hamidah Ibrahim.

Special gratitude to my family, and especially my beloved wife; Nan Zalina Long Abd. Wahab, my five children; Siti Fatimah Zaharah, siti Aisyah, Asma', Luqman, and Naemah, my mother; Chik Omar, my father; Mat Deris Muda, my mother in-law; Zaharah Yusof, for their patience and morale support.

I certify that an Examination Committee met on 4th December 2001 to conduct the final examination of Mustafa Mat Deris on his Doctor of Philosophy thesis entitled "Efficient Access of Replicated Data in Distributed Database Systems" in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulations 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

Ramlan Mahmod, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

Ali Mamat, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia
(Member)

Md. Yazid Mohd Saman, Ph.D.
Associate Professor
Kolej Universiti Sains dan Teknologi Malaysia
Mengabang, Telipot,
Kuala Terengganu
(Member)

Hamidah Ibrahim, Ph.D.
Associate Professor
Faculty of Computer Science and Information Technology,
Universiti Putra Malaysia
(Member)

Abu Talib Othman, Ph.D.
Professor /Dean of Information Technology School
Universiti Utara Malaysia
Sintok, Kedah
(Independent Examiner)

AINI IDERIS, Ph.D.
Professor,
Dean of Graduate School,
Universiti Putra Malaysia
Date: 10 JAN 2002

The thesis submitted to the Senate of Universiti Putra Malaysia has been accepted as fulfilment of the requirement for the degree of Doctor of Philosophy.

AINI IDERIS, Ph.D.
Professor
Dean of Graduate School
Universiti Putra Malaysia

Date: 14 MAR 2002

DECLARATION

I hereby declare that the thesis is based on my original work except for the citations, which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any degree at Universiti Putra Malaysia, UPM or at any other institution.

MUSTAFA MAT DERIS

Date:

TABLE OF CONTENTS

		Page
DEDICATION		ii
ABSTRACT		iii
ABSTRAK		v
ACKNOWLWDGEMENTS		vii
APPROVAL SHEETS		viii
DECLARATION FORM		x
LIST OF TABLES		xiv
LIST OF FIGURES		xv
LIST OF ABBREVIATIONS		xvi

CHAPTER

1	INTRODUCTION		1
	1.1	Data Replication	2
	1.2	Problem Statements	3
	1.3	Objectives and Scope of Research	5
	1.4	Organization of Thesis	6
2	REVIEW OF REPLICA CONTROL PROTOCOLS		7
	2.1	Read-One Write-All (ROWA)	7
	2.2	Voting (VT)	9
	2.3	Tree Quorum (TQ)	12
	2.4	Grid Configuration (GC)	15
3	FUNDAMENTAL CONCEPTS AND THEORY		21
	3.1	Formalization of Transaction Concept	21
	3.2	Serializability For Replicated Data	22
	3.3	Communication Cost	28
	3.4	Reliability and Availability	28
	3.5	Quorum Intersection Property	31

4 PERFORMANCE ANALYSIS OF EXIXTING DATA REPLICATION PROTOCOLS 33

4.1 Communication Cost Analysis 33
4.1.1 Read-One Write-All 33
4.1.2 Voting Protocol 34
4.1.3 Three Quorum Protocol 35
4.1.4 Grid Configuration Protocol 37
4.2 Operation Availability Model 37
4.3 Availability Analysis 38
4.3.1 Read-One Write-All 38
4.3.2 Voting Protocol 39
4.3.3 Three Quorum Protocol 40
4.3.4 Grid Configuration Protocol 41

5 THREE DIMENSIONAL GRID STRUCTURE MODEL.. 43
5.0 The Model 43
5.1 Three Dimensional Grid Structure Protocol 44
5.2 Correctness 48
5.3 Adding New Replica(s) 50
5.4 Deleting Current Replica(s) 51
5.5 Performance Analysis 53
5.5.1 Communication Cost 53
5.5.2 Availability 54
5.6 Performance Comparisons with other Protocols 59
5.6.1 Comparison of Communication Costs 59
5.6.2 Comparison of Availabilities 60

6 MANAGING REMOTE PROCEDURE CALL FOR TDGS 72
6.1 Remote Procedure Call 72
6.2 Three Dimensional Grid Structure Remote Procedure (TDGS-RP) Model 74
6.3 TDGS-RP Transaction Model 75
6.4 TDGS-RP Transaction Management 77
6.4.1 TDGS-RP Transaction Processing Model ... 77
6.4.2 Transaction Manager 78
6.4.3 Coordinating Algorithm for the Primary Replica 82
6.4.4 Cooperating Algorithm for the TDGS Replicas 86
6.4.5 Correctness 90
6.4.6 Examples 92

7 CONCLUSIONS AND FUTURE WORKS 97
7.1 Conclusions 97
7.2 Future Works 98

REFERENCES 100
APPENDIX 103
BIODATA OF THE AUTHORS 110

LIST OF TABLES

Table		Page
5.6.1	The read communication costs when n = 13, 40, and 121.....	60
5.6.2	The read availability when n= 13	66
5.6.3	The write availability when n = 13.....	66
5.6.4	The read availability when n= 40	69
5.6.5	The write availability when n = 40.....	69
6.4.6	Edges of replicas of each plane.....	93

LIST OF FIGURES

FIGURE		Page
2.3	A tree organization of 13 copies of a data object ..	14
2.4.1	A grid organization with 25 copies of a data object..	16
2.4.2	A grid organization with 25 copies of a data object and 11 of them are unavailable.....	17
5.1	A TDGS organization with 24 copies of a data object.....	47
5.2	The formation of TDGS logical structure when new copies are to be added.....	52
5.5.2	Two planes α and α consists of 3 x 3 copies of each plane.....	56
5.6.1	Comparison of the read cost between ROWA, VT, TQ, GC, and TDGS.....	62
5.6.2	Comparison of the read cost between ROWA, VT, TQ, GC, and TDGS.....	63
5.6.3	Comparison of the read availability between TDGS,GC,TQ,VT, and ROWA when n=13.....	67
5.6.4	Comparison of the write availability between TDGS,GC,TQ,VT, and ROWA when n =13.....	68
5.6.5	Comparison of the read availability between TDGS,GC,TQ,VT, and ROWA when n=40.....	70
5.6.6	Comparison of the write availability between TDGS,GC,TQ,VT, and ROWA when n=40.....	71
6.4.1	TDGS-RP transaction processing.....	80
6.4.6	A TDGS organization with 17 replicas.....	92
6.4.7	An example of TDGS-RP transaction processing...	95
6.4.8	The network is partitioned into two parts.....	96

LIST OF ABBREVIATIONS

GC	Grid Configuration
ROWA	Read-one Write-all
1SR	One-copy Serializable
TQ	Tree Quorum
TDGS	Three Dimensional Grid Structure
TDGS-RP	Three Dimensional Grid Structure Remote Procedure
VT	Voting

CHAPTER 1

INTRODUCTION

In early 1998, several research articles regarding distributed databases were published. Among them were those by Agrawal & Abbadi [1,2], Berstein et. al. [8], Chung [12], and Garcia-Molina & Barbara [16]. The articles revealed that data replication management is one of the current issues in distributed databases that has yet to be solved. It was on this basis that this study was initiated.

Distributed database system technology is one of the major recent developments in the database area, where it moves from centralization which resulted in monolithic gigantic databases towards more decentralization and autonomy of processing [15]. Many of today's commercial database systems such as *Oracle 8* and *IBM DB2 Propagator* provide the required support for data distribution and inter-database communication [27]. As new communication technologies are emerging, wireless and mobile computing concepts become reality and allow for even higher degrees of "distributedness" and flexibility in distributed databases.

With advances in distributed processing and distributed computing that occurred in the operating systems arena, the database research community did

considerable work to address the issues of data distribution, distributed query processing, distributed transactions management, and etc [13]. One of the major issue in data distribution is replicated data management. Typical replicated data management parameters are data availability and communication costs: the higher the data availability with lower communication costs the better the system is.

1.1 Data Replication

Replication is the act or result of reproducing- in short, a copy. As such, any type of data processing object can be implemented. Note that the definition describes replication as the act of reproducing. Therefore replication is much more than simply the copying of any object; it must also address the management of the complete copying process [10]. Thus, data replication is much more than simply copying data between data stores. It encompasses the administration and monitoring of a service that guarantees data consistency across multiple sites in a distributed environment.

In this evolving world of distributed databases, data replication plays an increasingly important role. It is a useful technique for distributed database systems where an object will be accessed (i.e., read and written) from multiple locations such as from a local area network environment or geographically

distributed world wide. For example student's results, will be read and updated by lecturers of various departments. Financial instruments' prices will be read and updated from all over the world [35]. This technique is used to provide high availability, fault tolerance, and enhanced performance [37]

The most common approaches to specify replication are synchronous and asynchronous replications. Synchronous means move or operate together at the same speed and in exact time with each other while asynchronous is otherwise. Thus, synchronous replication provides what is called 'tight consistency' between data stores. This means that the latency between data consistency is zero. Data at all sites/replicas is always the same, no matter from which replica the updated originated. While asynchronous replication provides what is called 'loose consistency' between data stores. This means that the latency between data consistency is always greater than zero; the replication process occurs asynchronously to the originating transaction. In other words, there is always some degree of lag between when the originating transaction is committed and when the effects of the transaction are available at any replica(s) [10].

1.2 Problem Statements

By storing multiple copies of data at several sites in the system, there is an increased data availability and accessibility to users despite site and

communication failures. It is an important mechanism because it enables organizations to provide users with access to current data where and when they need it. However, expensive synchronization mechanisms are needed to maintain the consistency and integrity of data [1]. This suggests that proper strategies are needed in managing replicated data.

One of the simplest protocols for managing replicated data is where read operations on an object are allowed to read any copy, and write operations are required to write all copies of the object. This protocol is termed as Read-One Write-All (ROWA) protocol. The ROWA protocol provides read operations with high degree of availability at low cost but severely restricts the availability of write operations since they cannot be executed at the failure of any copy. This protocol results in the imbalance of availability as well as the communication cost of read and write operations where read operations have a high availability and low communication cost whereas write operations have a low availability with higher communication cost. Voting protocols [6,12,13] became popular because they are flexible and easily implemented. One weakness of these protocols is that writing an object is fairly expensive: A write quorum of copies must be larger than the majority of votes. To optimize the communication cost and the availability of both read and write operations, the quorum protocol generalize the ROWA protocol by imposing the intersection requirement between read and write operations. Write operations can be made

fault-tolerant since they do not need to access all copies of objects. Dynamic quorum protocols have also been proposed to further increase availability in replicated databases [5,22]. However, these approaches do not address the issue of low-cost read operations. Tree quorum algorithm [2,3] uses quorums that are obtained from a logical tree structure imposed on data copies. However, this protocol has some drawbacks. If more than a majority of the copies in any level of the tree become unavailable, write execution cannot be executed.

Recently, several researchers [1,3,23,30] have proposed imposing logical structure on the set of copies in the database, and using logical information to create intersecting quorums. Protocol that use logical structure such as grid protocol, executes operation with low communication costs while providing fault-tolerance for both read and write operations. However, this protocol still requires that a bigger number of copies be made available to construct a quorum [22].

1.3 Objectives and Scope of Research

The objective of this research is on modeling a technique to optimize the communication costs and the availability of both read and write operations of replicated data in the distributed database systems. This thesis only concentrate on the implementation of replication based on zero latency before data

consistency is achieved, that is synchronous replication. Thus, this research will focus on replication protocol model in order to get high data availability with low communication costs in managing data replication by means of synchronous replication.

In this research, we describe the Three Dimensional Grid Structure (TDGS) protocol to improve read and write operations with respect to the logical structure. Consequently, the application of this protocol to Remote Procedure Call (RPC) through the combination of the TDGS replication and transaction management, will also be described.

1.4 Organization of Thesis

This thesis is organized as follows: In Chapter 2, we review five major replica control protocols; Read-One Write-All, Lazy Replication, Voting, Tree Quorum and Grid Configuration protocols. In Chapter 3, the concepts of transaction, serializability, communication cost, availability, and quorum intersection property that can be used in our study will be discussed. In Chapter 4, the performance analysis of the existing replication protocols are given. In Chapter 5, the model of the TDGS and performance comparisons with other models are presented. In Chapter 6, managing remote procedure call for TDGS protocol is presented. In Chapter 7, the conclusion and the future works will be presented.

CHAPTER 2

REVIEW OF REPLICA CONTROL PROTOCOL

This chapter reviews some of the major replica control protocols namely the ROWA protocol, Lazy replication protocol, Voting (VT) protocol, Tree Quorum (TQ) protocol, and Grid Configuration (GC) protocol. These protocols are then compared to the proposed TDGS protocol, particularly in terms of data availability and communication cost.

2.1 Read-One Write-All Protocol

The simplest technique to maintain replicated data is that a read operation is allowed to read any copy, and a write operation is required to write all copies of data object which is called Read-One Write-All (ROWA) protocol. This protocol works correctly since a transaction processes from one correct state to another correct state. The ROWA has the lowest read cost because only one replica is accessed by a read operation. The weakness of this method is the low write availability because a write operation cannot be done at the failure of any replica.

The available copies technique proposed by [8] is an enhanced version of the ROWA approach in terms of the availability of write operations. Every read is

translated into a read of any replica of the data object and every write is translated into write of all available copies of that data object. This technique handles each site either it is operational or down and that all operational sites can communicate with each other. Therefore, each operational site can independently determine which sites are down, simply by attempting to communicate with them. If a site does not respond to a message within the timeout period, then it is assume to be down. However, writing is very expensive when all copies are available: forcing read-write transactions to write all replicas.

Lazy replication protocol do not attempt to perform the write operation on all copies of the data object within the context of the transaction that updates that data object. Instead, they perform the update on one or more copies of the data object and later propagate the changes to all the other copies in all sites. A lazy replication scheme can be characterized using four basic parameters [9,18,27]: The *ownership* parameter defines the permissions for updating copies. If a copy is updatable it is called *primary* copy, otherwise it is called a *secondary* copy. The site that stores the primary copy of an data object is called a master for this data object, while the sites that store its secondary copies are called *slaves*. The *propagation* parameter defines when the updates copy must be propagated towards the sites storing the other copies of the same data object. Generally, lazy replication protocols can be classified into two groups. The first group consists